Please Click here to view the drawing

Korean FullDoc. English Fulltext

(19) KOREAN INTELLECTUAL PROPERTY OFFICE

KOREAN PATENT ABSTRACTS

(11)Publication number: 1020010105241 A
(43)Date of publication of application: 28.11.2001

(21)Application number: 1020010026934
(22)Date of filing: 17.05.2001

(71)Applicant: MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.

(72)Inventor: ARAKI SHOICHI
KUTSUMI HIROSHI
MARUNO SUSUMU
NAITO EIICHI
OZAWA JUN

(51)Int. Cl G06F 17/30

(54) INFORMATION RETRIEVAL SYSTEM

(57) Abstract:

PURPOSE: An information retrieval system is provided to calculate feature vectors of documents, to classify the documents into clusters based on the calculated feature vectors for displaying document retrieval results together for each cluster, to retrieve a question similar to a user question upon receipt of a question from the user, and to present an answer associated with the retrieved question so that it can reduce the burden on the user in an information retrieval and easily update information in an information retrieval system. CONSTITUTION: The information retrieval system comprises a document storage section(11), a cluster storage section(12), a cluster label storage section(13), a document label storage section(14), a feature vector extraction section(15), a clustering section(16), a cluster label preparation section(17), a document label preparation section(18), a database retrieval section(19), an interface section (20), a user input section(21), and a user display section(22). The document storage section(11) stores a plurality of documents. The feature vector extraction section(15) extracts feature vectors from the documents stored in the document storage section(11). The clustering section 16 classifies the documents stored in the document storage section(11) into clusters based on the feature vectors extracted by the feature vector extraction section(15). The cluster storage section(12) stores the clusters into which the clustering section(16) has classified the documents. The cluster label preparation section(17) prepares cluster labels representing the contents of the respective clusters created by the clustering section(16). Each cluster label may be a term label composed of a term or may be a sentence label composed of a sentence. The cluster label storage section(13) stores the cluster labels prepared by the cluster label preparation section(17). The document label preparation section(18) prepares document labels representing the contents of the documents as elements of the respective clusters created by the clustering section(16). The document label storage section(14) stores the document labels prepared by the document label preparation section(18). The interface section(20) manages input/output with the user. The database retrieval section(9) retrieves a document that satisfies the retrieval condition, from the document storage section(11). The user display section(22) displays the retrieval results for the user.

(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(51) Int. Cl.
G06F 17/30

(11) 공개번호 특2001-0105241
(43) 공개일자 2001년11월28일

(21) 출원번호 10-2001-0026934
(22) 출원일자 2001년05월17일
(30) 우선권주장 특원2000-145168 2000년05월17일 일본(JP)
(71) 출원인 마츠시타 덴끼 산교 가부시키가이샤
일본 오오사카후 가도마시 오오아자 가도마 1006
(72) 발명자 나이토에이이치
일본국교토후교타나베시다키기구와노카2-31
아라키쇼이치
일본국오오사카후오오사카시조토쿠이마후쿠히가시3-15-22-503
구초미히로시
일본국오오사카후모리구치시데라카타니시키도리1-4-31
오자와준
일본국나라켄나라시오부치쵸3810-2-506
마루노스스무
일본국교토후교타나베시야마테미나미4-4-3
(74) 대리인 김영철

심사청구 : 없음

(54) 정보검색 시스템

요약

본 발명은 정보검색에 소요되는 이용자의 부담을 줄이고 검색대상정보의 자동갱신을 실현하기 위한 정보검색 시스템에 관한 것이다. 주요 구성으로서, 서로 대응된 질문과 회답을 문서기억부(11)에 기억하고, 클러스터 분류부(16)는 문서기억부(11) 중의 각 회답의 특징벡터에 기초하여 회답을 클러스터 분류하고, 데이터베이스 검색갱신부(33)는 이용자로부터 자유문에 의한 질문이 입력된 경우에 유사질문을 검색하여 대응하는 회답을 클러스터별로 정리하여, 이용자 또는 전문가에게 제시하고, 데이터베이스 검색갱신부(33)는 이용자 또는 전문가가 가장 적절하다고 판단한 회답을 선택하였을 때에는 그 선택된 회답을 기초로 적절한 회답이 없을 경우 전문가가 새로 입력한 회답을 기초로 문서기억부(11)를 자동갱신함으로써 전문가가 입력한 자유문에 의한 회답이 그대로 이용자에게 제시되는 것을 특징으로 한다.

대표도

도14

색인어

문서기억부, 클러스터 분류부, 데이터베이스 검색갱신부

명세서

도면의 간단한 설명

도 1은 본 발명의 제 1 실시예에 관한 정보검색 시스템의 구성을 도시한 블록도

도 2는 도 1 중의 문서기억부에 기억된 문서의 예를 도시한 도면

도 3은 도 1 중의 이용자 표시부에서의 검색결과의 표시예를 도시한 도면

도 4는 도 1 중의 특징벡터 추출부의 처리순서를 도시한 플로우차트

도 5는 추출된 문서특징벡터의 예를 도시한 도면

도 6은 도 1 중의 클러스터 분류부의 처리순서를 도시한 플로우차트

도 7은 클러스터 분류결과의 예를 도시한 도면

도 8은 도 1 중의 클러스터 라벨작성부에서의 단어라벨작성순서를 도시한 플로우차트

도 9는 작성된 단어라벨의 예를 도시한 도면

도 10은 도 1 중의 클러스터 라벨작성부에서의 문서라벨 작성순서를 도시한 플로우차트

도 11은 작성된 문서라벨의 예를 도시한 도면

도 12는 도 1 중의 문서라벨 작성부의 처리순서를 도시한 플로우차트

도 13은 작성된 문서라벨의 예를 도시한 도면

도 14는 본 발명의 제 2 실시예에 관한 정보검색 시스템의 구성을 도시한 블록도

도 15는 도 14 중의 문서기억부에 기억된 문서 중의 질문표 부분의 예를 도시한 도면

도 16은 도 14 중의 문서기억부에 기억되어 있는 문서중의 회답표 부분의 예를 도시한 도면

도 17은 도 14 중의 전문가 표시부에서의 검색결과의 표시예를 도시한 도면

도 18은 도 14 중의 이용자 표시부에서의 검색결과의 표시예를 도시한 도면

도 19는 도 14 중의 특징벡터 추출부에서의 이용자 질문의 특징벡터 추출순서를 도시한 플로우차트

도 20은 이용자 질문으로부터 추출된 특징벡터의 예를 도시한 도면

도 21은 도 14 중의 유사도 연산부의 처리순서를 도시한 플로우차트

도 22는 도 14 중의 데이터베이스 검색갱신부의 처리순서를 중심으로 도시한 플로우차트

* 도면의 주요 부분에 대한 부호의 설명 *

11 : 문서기억부	12 : 클러스터 기억부
13 : 클러스터라벨 기억부	14 : 문서라벨 기억부
15 : 특징벡터 추출부	16 : 클러스터 분류부
17 : 클러스터라벨 작성부	18 : 문서라벨 작성부
19 : 데이터베이스 검색부	20 : 인터페이스부
21 : 이용자 입력부	22 : 이용자 표시부
31 : 특징벡터 기억부	32 : 유사도 연산부
33 : 데이터베이스 검색갱신부	41 : 전문가 입력부
42 : 전문가 표시부	

발명의 상세한 설명

발명의 목적

발명이 속하는 기술분야 및 그 분야의 종래기술

본 발명은 대량의 정보 중에서 이용자가 구하는 정보를 용이하게 찾을 수 있게 하기 위한 정보검색 시스템에 관한 것이다.

최근 인터넷의 보급에 따라 WWW(World Wide Web) 상에 HTML(Hyper Text Markup Language)로 기술된 여러 가지 홈페이지가 게재되는 등으로 인해 일반 이용자는 대량의 정보를 액세스할 수 있다. 또 빈번히 문의되는 질문과 그 회답을 쌍으로 한 FAQ(Frequently Asked Questions) 모음집이라는 리스트가 공개되어 있어 이용자는 질문에 대한 회답을 얻을 수 있다. 이들 정보는 이용자가 구하는 정보의 소재를 알면 바로 열람할 수 있으므로 편리하지만 반대로 대량의 정보 중에서 자기가 구하는 정보를 찾는 것은 대단한 작업이 되고 있다.

이 때문에 문서로부터 키워드를 잘라내어 그 문서의 특징량으로 하고, 특징량 사이의 내적을 산출하고 문서간의 유사도를 구하여 질문에 대한 유사문서를 검색하는 검색기술이 알려져 있다.

발명이 이루고자 하는 기술적 과제

그러나 인터넷 상의 정보나 또는 사례를 기초로 축적된 FAQ 모음집은 많은 사람이 독립적으로 정보를 제공하고 있으므로 정보의 중복을 피할 수 없어 같은 내용의 문서가 대량 존재한다. 따라서 종래의 기술에서는 질문문에 유사한 문서로서 같은 내용의 문서가 대량 검색되는 일이 많으므로 이용자는 결국 대량의 검색결과 중에서 원하는 정보를 찾는 작업이 필요하였다. 검색결과를 일정한 수로 제한하면 자기가 원하는 정보가 없거나 하는 문제점이 있었다.

또 이용자가 검색결과로부터 원하는 정보를 찾는데 성공하더라도 그것이 FAQ 모음집에 반영되지 않으므로 다른 이용자가 같은 조건으로 검색한 경우에도 마찬가지로 찾는 수고가 필요하였다. 정보의 중복을 피하면서 FAQ 모음집을 보다 충실하게 하려면 같은 정보가 이미 존재하고 있는지의 여부를 체크해야 하므로 정보제공자에게 부담이 되고 있었다.

본 발명의 목적은 이용자의 정보검색에 소요되는 부담을 줄이는 정보검색 시스템을 제공하는 것에 있다.

본 발명의 다른 목적은 검색대상의 정보를 용이하게 갱신할 수 있는 정보검색 시스템을 제공하는 것에 있다.

발명의 구성 및 작용

상기 목적을 달성하기 위해 본 발명은 문서의 특징벡터를 산출하고 특징벡터에 기초하여 문서를 클러스터 분류하고, 문서의 검색결과를 클러스터별로 정리하여 표시하는 것이다. 이로 인하여 이용자는 검색결과를 유사한 문서의 집합으로서 파악하기가 용이해 진다.

또 본 발명은 이용자로부터의 질문이 입력된 경우에 유사질문을 검색하여 대응하는 회답을 이용자 또는 전문가에게 제시하고 이용자 또는 전문가가 가장 적절하다고 판단한 회답을 선택하였을 때 그 선택된 회답을 기초로 문서 데이터베이스를 자동적으로 갱신하는 것이다. 적절한 회답이 없을 때에는 전문가가 새롭게 입력한 회답을 기초로 문서 데이터베이스를 갱신한다. 이로 인하여 다음에 같은 질문이 입력된 경우에 적절한 회답을 할 수 있다.

(실시예)

이하 본 발명의 두가지 실시예에 대하여 도면을 참조하여 설명한다.

(제 1 실시예)

도 1은 본 발명의 제 1 실시예에 관한 정보검색 시스템의 구성을 도시한다. 도 1의 정보검색 시스템은 문서기억부(11), 클러스터 기억부(12), 클러스터라벨 기억부(13), 문서라벨 기억부(14), 특징벡터 추출부(15), 클러스터 분류부(16), 클러스터 라벨작성부(17), 문서라벨 작성부(18), 데이터베이스 검색부(19), 인터페이스부(20), 이용자 입력부(21) 및 이용자 표시부(22)로 구성되고, 예를 들어 인터넷을 통해 서로 접속된 문서서버와 이용자단말로 실현된다. 문서기억부(11)는 복수의 문서를 기억한다. 특징벡터 추출부(15)는 문서기억부(11)에 기억된 문서로부터 특징벡터를 추출한다. 클러스터 분류부(16)는 특징벡터 추출부(15)가 구한 특징벡터에 기초하여 문서기억부(11)에 기억된 문서를 클러스터 분류한다. 클러스터 기억부(12)는 클러스터 분류부(16)가 클러스터 분류한 문서의 클러스터를 기억한다. 클러스터 라벨작성부(17)는 클러스터 분류부(16)가 작성한 각 클러스터에 대하여 당해 클러스터의 내용을 나타내는 클러스터 라벨을 작성한다. 클러스터 라벨은 단어로 이루어지는 단어라벨 또는 문장으로 이루어지는 문장라벨이다. 클러스터 라벨기억부(13)는 클러스터 라벨작성부(17)가 작성한 클러스터 라벨을 기억한다. 문서라벨 작성부(18)는 클러스터 분류부(16)가 작성한 클러스터의 요소인 각 문서에 대하여 그 문서의 내용을 나타내는 문서라벨을 작성한다. 문서라벨 기억부(14)는 문서라벨 작성부(18)가 작성한 문서라벨을 기억한다. 이용자 입력부(21)는 이용자로부터 주어진 검색조건을 접수한다. 검색조건으로서는 문서의 키워드, 문서 ID 등 문서검색의 조건이 된다면 무엇이든지 좋다. 인터페이스부(20)는 이용자와의 입출력을 관리한다. 데이터베이스 검색부(19)는 문서기억부(11)로부터 검색조건을 만족하는 문서를 검색한다. 이용자 표시부(22)는 검색결과를 이용자에게 제시한다.

도 2는 도 1 중 문서기억부(11)에 기억된 문서의 예를 도시한다. 문서기억부(11)에는 검색의 대상이 되는 주어진 n(n ≥2)개의 문서가 기억된다. 각 문서는 독특한 문서 ID와 문장형식의 본문으로 이루어진다. i번째의 문서를 Di라 한다(1 ≤i ≤n).

도 3은 도 1 중의 이용자 표시부(22)에서의 검색결과의 표시예를 도시한다. 도 3에 의하면 어떤 검색조건에 대한 문서의 검색결과가 클러스터별로 정리되어 표시된다. 구체적으로는 클러스터 ID와 그 클러스터에 포함되는 문서의 문서 ID 및 본문을 클러스터별로 표형식으로 표시하고, 마우스로 「이전 클러스터」 버튼이나 「다음 클러스터」 버튼을 눌러 다른 클러스터를 표시함으로써 모든 검색결과를 표시할 수 있다. 이로 인하여 이용자는 검색결과를 유사한 문서의 집합으로서 파악하는 것이 용이해 진다. 더구나 표시된 클러스터에는 당해 클러스터의 내용을 나타내는 클러스터 라벨이 표시되는 동시에 문서라벨에 지정된 문장이 밑줄로 표시된다. 따라서 클러스터의 내용을 이용자가 파악하기 쉽게 된다. 또 검색결과로서 클러스터 ID, 문서 ID까지도 표시하였지만 표시하지 않아도 된다.

이하 상기 제 1 실시예의 상세한 사항을 문서등록시의 동작과 문서검색시의 동작으로 나누어 설명한다. 문서등록시의 동작이란 처음으로 문서가 문서기억부(11)에 등록되는 경우 또는 그 이후에 문서의 추가/변경/삭제가 있는 경우의 동작이다. 문서검색시 동작이란 등록문서를 검색하여 열람하는 경우의 동작이다.

(문서등록시의 동작)

도 4는 도 1 중의 특징벡터 추출부(15)의 처리순서를 도시한다. 우선 특징벡터 추출부(15)는 문서기억부(11)에 기억된 모든 문서 Di를 차례로 인출하고, 각 문서 Di의 특징벡터 Vi를 추출한다. 특징벡터는 문서의 특징을 나타내는 단어 Tj와 그 가중값 Wij의 그룹을 요소로 하는 벡터이고, 그 요소의 수는 문서에 따라 다르다. 여기에서 j는 단어를 식별하는 독특한 번호이다. 도 4에서 단계 S101에서는 문서의 카운터 i에 i=1을 설정한다. 단계 S102에서는 문서기억부(11)로부터 문서 Di를 인출하여 형태소 해석, 구문 해석, 불필요어 제거 등 일반적으로 알려진 방법으로 출현하는 단어 Tj를 본문에서 추출하여 문서 Di 내에서의 단어 Tj의 출현횟수 Fij를 카운트한다. 종료판정단계 S103에서는 전체 문서에 대해 단계 S102의 처리가 종료된 경우, 즉 i=n인 경우에는 S105로 진행한다. 그렇지 않은 경우에는 S104로 진행한다. 단계 S104에서는 카운터 i를 1 증가하여 단계 S102로 진행한다. 단계 S105에서는 단어 Tj의 전체 문서에 대한 중요도로서 단어 Tj가 출현하는 문서수의 적음을 나타내는 정도, 즉 IDF(inverse document frequency)값을 수학식 1로 산출한다.

$$IDFj = \log \frac{n}{Mj} + 1$$

여기에서 Mj는 단어 Tj가 출현하는 문서의 수를 나타낸다. 단계 S106에서는 문서의 카운터 i에 i=1을 설정한다. 단계 S107에서는 단어 Tj가 문서 Di를 특징짓는 가중값 Wij로서 문서 Di 내에서의 단어 Tj의 출현비율을 나타내는 TF(term frequency)값과 상기 IDF값을 곱한 TFIDF값을 수학식 2로 산출한다.

$$Wij = \frac{Fij}{\sum_{j:Tj \in Di} Fij} \cdot IDFj$$

종료판정단계 S108에서는 전체 문서에 대하여 단계 S107의 처리가 종료된 경우, 즉 i=n인 경우에는 종료한다. 그렇지 않은 경우에는 S109로 진행한다. 단계 S109에서는 카운터 i를 1 증가하여 단계 S107로 진행한다.

도 5는 추출된 문서특징벡터 Vi의 예를 도시한다. 또 상기 특징벡터의 산출에서는 TFIDF값을 이용하고 있었지만 단순히 단어의 출현횟수로 하는 등 다른 방법으로 해도 된다.

도 6은 도 1 중의 클러스터 분류부(16)의 처리순서를 도시한다. 클러스터 분류부(16)는 특징벡터 추출부(15)가 추출한 특징벡터를 이용하여 모든 문서를 m개의 클러스터로 분류한다(1 < m < n). 여기에서 k번째의 클러스터를 Ck라 한다(1 ≤k ≤m). 클러스터 분류의 순서로서 트리형으로 점차 클러스터에 분류해 가는 계층적 클러스터링을 이용한다. 도 6에서 단계 S111에서는 클러스터간 거리의 초기계산을 한다. 여기에서는 초기 클러스터로서 각각 1개의 문서 Di만을 요소로서 갖는 n개의 클러스터 Ci를 설정하거나 각 클러스터 Ck, Cl(1 ≤k, l ≤n) 사이의 거리 Lkl로서 각 문서의 특징벡터간의 거리를 나타내는 수학식 3의 유사비를 채용한다.

$$Lkl = -\log \frac{\sum_{j:Tj \in Dk \cup Dl} MIN(Wkj, Wlj)}{\sum_{j:Tj \in Dk \cup Dl} MAX(Wkj, Wlj)}$$

단계 S112에서는 클러스터링 횟수의 카운터 i에 i=1을 설정한다. 단계 S113에서는 모든 클러스터의 조합중에서 클러스터간 거리 Lkl이 가장 작은 클러스터 Ck, Cl(k<l)의 그룹을 탐색한다. 단계 S114에서는 클러스터 Ck, Cl을 통합하여 클러스터 Cg라 한다. 즉 Cg=Ck∪Cl, Cl=∅ 로 한다(∅는 공집합을 나타낸다). 클러스터의 통합에 따라 클러스터 Cg와 다른 클러스터 Ch(1 ≤h ≤n)의 클러스터간 거리를 워드(woad)법을 이용하여 수학식 4로 산출한다.

$$Lgh = \frac{(Nk + Nh) \cdot Lkh + (Nl + Nh) \cdot Llh - Nh \cdot Lkl}{Ng + Nh}$$

여기에서 Nk는 클러스터 Ck의 요소의 수이다. 종료판정단계 S115에서는 클러스터링 횟수가 n-1인 경우, 즉 모든 초기 클러스터가 1개의 클러스터에 통합된 경우에는 단계 S117로 진행한다. 그렇지 않은 경우에는 S116으로 진행한다. 단계 S116에서는 카운터 i를 1 증가하여 단계 S112로 진행한다. 단계 S117에서는 클러스터 수를 결정한다. 단계 S111부터 단계 S115까지의 클러스터 분류과정에서는 클러스터링 횟수마다 클러스터의 수는 하나씩 감소한다. 단계 S117에서는 클러스터 분류과정을 되돌아보아 적절한 클러스터링 횟수를 결정한다. 여기에서는 요소를 2개 이상 갖는 클러스터의 수가 최대가 되는 클러스터링 횟수를 적절한 클러스터링 횟수라 한다. 단계 S118에서는 단계 S117에서 결정한 클러스터링 횟수까지 클러스터 분류를 한 시점에서의 각 클러스터에 포함되는 요소를 클러스터 기억부(12)에 기입한다.

도 7은 클러스터 기억부(12)에 기입된 클러스터의 예를 도시한다. 각 클러스터는 클러스터 ID와 그 클러스터에 포함되는 문서의 문서 ID로 이루어진다. 예를 들면 클러스터 1에는 1, 190, 432, 644번의 4개의 문서가 포함된다. 이것은 이들 4개의 문서의 특징벡터끼리가 다른 문서에 비해 유사한 것을 나타낸다. 또 상기의 예에서는 클러스터 분류방법으로서 계층적 클러스터링을 이용하였지만 비계층적 클러스터링이라도 된다. 초기클러스터간 거리로서 수학식 3의 유사비를 이용하였지만 유클리드 평방거리 등 다른 거리를 이용해도 된다. 클러스터 통합시의 클러스터간 거리의 산출방법으로서 수학식 4의 워드법을 이용하였지만, 최장거리법 등 다른 방법을 이용해도 된다. 클러스터수의 결정방법으로서 요소를 2개 이상 갖는 클러스터의 수가 최대가 되는 클러스터링 횟수로 하였지만 클러스터 수를 문서수의 일정한 비율로 하는 등 다른

결정방법이라도 된다.

도 8은 도 1 중의 클러스터라벨 작성부(17)에서의 단어라벨 작성순서를 도시한다. 단계 S201에서는 클러스터의 카운터 k에 k=1을 설정한다. 단계 S202에서는 클러스터 Ck의 요소인 모든 문서 Di의 특징벡터 Vi에 포함되는 단어 Tj마다 클러스터 Ck의 요소인 문서 Di 중 단어 Tj가 출현하는 출현문서수를 카운트한다. 단계 S203에서는 클러스터 Ck의 요소인 모든 문서 Di에 포함되는 단어 Tj마다 단어 Tj의 TFIDF값(=wij)의 클러스터 Ck의 요소인 모든 문서 Di에 대한 합계를 산출한다. 단계 S204에서는 클러스터 Ck의 요소인 모든 문서 Di의 특징벡터 Vi에 포함되는 모든 단어 Tj를 단계 S202에서 구한 출현문서수가 많은 순서로 분류한다. 출현문서수가 같은 경우는 단계 S203에서 구한 TFIDF값의 합계가 큰 차례로 분류한다. 단계 S205에서는 단계 S204에서 분류된 상위의 3개의 단어를 선택하고 클러스터의 단어라벨로서 클러스터라벨 기억부(13)에 기입한다. 종료판정단계 S206에서는 전체 클러스터에 대해 단계 S202부터 단계 S205까지의 처리가 종료된 경우, 즉 k=m인 경우에는 종료한다. 그렇지 않은 경우에는 S207로 진행한다. 단계 S207에서는 카운터 k를 1 증가하고 단계 S202로 진행한다.

도 9는 클러스터 라벨기억부(13)에 기입된 단어라벨의 예를 도시한다. 예를 들어 클러스터 1에는 「과자」, 「간식」 「치즈」라는 단어라벨이 붙어 있는 것을 나타낸다. 또 단어라벨의 작성방법으로서 단어의 출현문서수로 분류하였지만, TFIDF값만으로 분류하는 등 다른 방법이라도 된다. 또 단어라벨의 단어수를 3개로 하였으나 3개가 아니어도 된다.

도 10은 도 1 중의 클러스터라벨 작성부(17)에서의 문장라벨 작성순서를 도시한다. 단계 S301에서는 클러스터의 카운터 k에 k=1을 설정한다. 단계 S302에서는 클러스터 Ck의 요소인 모든 문서 Di의 특징벡터 Vi에 포함되는 단어 Tj마다 클러스터 Ck의 요소인 문서 Di 중 단어 Tj가 출현하는 출현문서수를 카운트한다. 단계 S303에서는 클러스터 Ck의 요소인 모든 문서 Di를 구성하는 문장별로 그 문장에 포함되는 단어 Tj의 합계, 즉 단계 S302에서 카운트한 출현문서수의 합계를 산출한다. 여기에서 문장이란 문서를 「, 」등의 구점으로 구분한 하나하나의 문자열을 말한다. 단계 S304에서는 클러스터 Ck의 요소인 모든 문서 Di를 구성하는 문장을 단계 S303에서 구한 출현문서수의 합계가 큰 순서로 분류한다. 단계 S305에서는 단계 S304에서 분류된 최상위의 문장을 선택하고 클러스터의 문장라벨로서 클러스터라벨 기억부(13)에 기입한다. 최상위의 문장이 복수 있는 경우에는 그 중에서 문자수가 최소인 문장을 선택한다. 종료판정단계 S306에서는 전체 클러스터에 대하여 단계 S302부터 단계 S305까지의 처리가 종료된 경우, 즉 k=m인 경우에는 종료한다. 그렇지 않은 경우에는 S307로 진행한다. 단계 S307에서는 카운터 k를 1 증가하여 단계 S302로 진행한다.

도 11은 클러스터라벨 기억부(13)에 기입된 문장라벨의 예를 도시한다. 예를 들어 클러스터 1에는 「수분이 많은 것(젤리, 푸딩, 요구르트)을 」이라는 문장라벨이 붙어 있는 것을 나타낸다. 또 문장라벨의 작성방법으로서 단어의 출현문서수의 합계로 분류하였으나 TFIDF값의 합계로 분류하는 등 다른 방법이라도 된다. 또 출현문서수의 합계가 최상위의 문장이 복수개 있는 경우에 문자수가 최소인 문장을 선택하였으나 문장의 개시위치가 가장 앞쪽인 문장을 선택하는 등 다른 방법으로 해도 된다.

도 12는 도 1 중의 문서라벨 작성부(18)의 처리순서를 도시한다. 단계 S401에서는 문서의 카운터 i에 i=1을 설정한다. 단계 S402에서는 문서 Di를 구성하는 각 문장마다 그 문장에 포함되는 모든 단어 Tj의 TFIDF값(=wij)의 합계를 산출한다. 종료판정단계 S403에서는 모든 문서에 대해 단계 S402의 처리가 종료된 경우, 즉 i=n인 경우에는 S405로 진행한다. 그렇지 않은 경우에는 S404로 진행한다. 단계 S404에서는 카운터 i를 1 증가하고 단계 S402로 진행한다. 단계 S405에서는 클러스터의 카운터 k에 k=1을 설정한다. 단계 S406에서는 클러스터 Ck의 요소인 모든 문서 Di를 구성하는 문장을 단계 S402에서 구한 합계가 많은 순서로 분류한다. 단계 S407에서는 문서 Di의 문서라벨로서 단계 S406에서 분류된 최상위의 문장을 선택한다. 단 선택된 문장이 클러스터 라벨작성부(17)가 작성한 클러스터의 문장라벨과 동일한 경우에는 문서 Di의 문서라벨로서 단계 S406에서 분류된 상위로부터 두번째의 문장을 선택한다. 단계 S408에서는 단계 S407에서 선택된 문서 Di의 문서라벨을 문서라벨 기억부(14)에 기입한다. 종료판정단계 S409에서는 전체 클러스터에 대해 단계 S406부터 단계 S408까지의 처리가 종료된 경우, 즉 k=m인 경우에는 종료한다. 그렇지 않은 경우에는 S410으로 진행한다. 단계 S410에서는 카운터 k를 1 증가하고 단계 S406으로 진행한다.

도 13은 문서라벨 기억부(14)에 기입한 문서라벨의 예를 도시한다. 예를 들어 클러스터 1에 포함되는 문서 1에는 「씹는 효과가 있고 미련이 남지 않는 것으로, 」라는 문서라벨이 붙어 있는 것을 나타낸다.

이상의 동작에 의해 문서등록시에 각 문서에 대하여 특징벡터를 추출하고 클러스터라벨 및 문서라벨을 작성하여 각각의 기억부에 기억해 둔다.

(문서검색시의 동작)

우선 인터페이스부(20)는 이용자 입력부(21)를 통해 문서의 검색조건을 접수한다. 데이터베이스 검색부(19)는 검색조건을 만족하는 문서를 문서기억부(11)로부터 검색하고 당해 검색된 문서가 포함되는 클러스터를 클러스터 기억부(12)로부터 검색하고 당해 검색된 클러스터에 포함되는 문서를 다시 문서기억부(11)에서 검색하여 그 결과를 클러스터라벨 및 문서라벨과 함께 인터페이스부(20)로 보낸다. 인터페이스부(20)는 이용자 표시부(22)를 통해 검색결과를 이용자에게 제시한다(도 3).

또 본 실시예에서는 주어진 것이 문서에 미리 기억되어 있었으나 광디스크 등의 기억매체나 인터넷 등의 네트워크 매체 등에 의해 뒤에서부터 새롭게 도입하거나 개정되어도 된다. 또 문서의 검색은 키워드나 문서 ID에 의한 것 이외에 전문검색이거나 연산자검색이어도 된다.

(제 2 실시예)

도 14는 본 발명의 제 2 실시예에 관한 정보검색 시스템의 구성을 도시한다. 도 14의 정보검색 시스템은 이용자의 자유문에 의한 질문에 대하여 과거의 사례검색에 기초하는 적절한 회답을 하는 시스템으로서, 예를 들어 인터넷을 통해 서로 접속된 문서 서버, 이용자 단말 및 전문가 단말로 실현된다. 도 14의 구성은 도 1의 구성에 특징벡터 기억부(31), 유사도 연산부(32), 전문가 입력부(41) 및 전문가 표시부(42)를

추가하며, 도 1 중의 데이터베이스 검색부(19)를 데이터베이스 검색갱신부(33)로 치환한 것이다. 문서기억부(11)는 서로 대응된 복수의 질문문서와 복수의 회답문서를 기억한다. 전문가 표시부(42)는 전문가에게 검색결과를 제시한다. 전문가 입력부(41)는 전문가로부터의 선택입력 및 자유문에 의한 회답입력을 접수한다. 인터페이스부(20)는 이용자 및 전문가와의 입출력을 관리한다. 특징벡터 추출부(15)는 문서기억부(11)의 질문문서 및 회답문서의 각각으로부터 특징벡터를 추출하는 기능과, 이용자의 자유문에 의한 질문입력으로부터 특징벡터를 추출하는 기능과, 전문가의 자유문에 의한 회답입력으로부터 특징벡터를 추출하는 기능을 갖는다. 특징벡터 기억부(31)는 특징벡터 추출부(15)가 문서기억부(11)의 질문문서 및 회답문서의 각각으로부터 추출한 특징벡터를 기억한다. 유사도 연산부(32)는 이용자 질문입력으로부터 추출된 특징벡터와, 특징벡터 기억부(31)가 기억하고 있는 질문문서의 특징벡터의 유사도를 구하는 기능과, 전문가 회답입력으로부터 추출된 특징벡터와 특징벡터 기억부(31)가 기억하고 있는 회답문서의 특징벡터의 유사도를 구하는 기능을 갖는다. 데이터베이스 검색갱신부(33)는 문서기억부(11)의 문서를 검색하는 기능에 덧붙여서 이용자 또는 전문가의 응답에 기초하여 문서기억부(11)를 갱신하는 기능을 갖는다.

도 15 및 도 16은 도 14 중의 문서기억부(11)에 기억된 문서의 예를 도시한다. 도 15는 질문문서를 모은 질문표의 부분을 나타낸다. 이 질문표는 독특한 질문 ID, 문장형식의 질문 및 그 질문에 대응하는 회답 ID로 이루어진다. 도 16은 회답문서를 모은 회답표의 부분을 나타낸다. 이 회답표는 독특한 회답 ID 및 문장형식의 회답으로 이루어진다. I번째의 질문을 Qi로 하고 K번째의 회답을 Ak로 한다($1 \leq i \leq n$ 이되 $1 \leq k \leq m$). 여기에서 $n \geq m$의 관계가 성립된다. 즉 복수의 질문에 대하여 1개의 회답이 대응하는 경우가 있다.

도 17은 도 14 중의 전문가 표시부(42)에서의 검색결과의 표시예를 도시한다. 도 17에서는 이용자로부터의 질문에 덧붙여서 회답 후보가 클러스터에 분류된 상태에서 클러스터의 문서라벨 및 클러스터 중의 문서라벨과 함께 표시된다. 도 17에서는 마우스로 「이전 페이지」 버튼이나 「다음 페이지」 버튼을 눌러 다른 페이지를 표시함으로써 모든 검색결과를 표시할 수 있다. 이로 인하여 전문가는 유사한 문서의 집합으로서 표시된 검색결과를 참조하여 가장 적절한 회답을 용이하게 선택할 수 있다. 또는 자유문에 의한 전문가 회답을 입력할 수도 있다. 또 도 17의 예에서는 클러스터 라벨로서 문서라벨을 표시하였으나 이것과 함께 또는 이것 대신에 단어라벨을 표시해도 된다. 또 검색결과로서 클러스터 ID, 문서 ID까지도 표시하였으나, 표시하지 않아도 된다.

도 18은 도 14 중의 이용자 표시부(22)에서의 검색결과의 표시예를 도시한다. 여기에서는 번호 1의 문서가 전문가회답으로서 선택된 것이다.

이하 상기 제 2 실시예의 상세한 내용을 제 1 실시예와 마찬가지로 문서등록시의 동작과 문서검색시의 동작으로 나누어 설명한다.

(문서등록시의 동작)

우선 특징벡터 추출부(15)는 문서기억부(11)에 기억된 모든 문서로부터 질문의 특징벡터 VQi와 회답의 특징벡터 VAk를 추출하여, 추출된 특징벡터를 특징벡터 기억부(31)에 기입한다. 특징벡터의 추출순서는 제 1 실시예와 같다. 제 1 실시예와의 차이는 질문과 회답부분에 대하여 각각 특징벡터를 산출하는 점과, 특징벡터를 특징벡터 기억부(31)에 기입하는 점이다.

다음으로 클러스터 분류부(16)는 특징벡터 기억부(31)로부터 회답의 특징벡터 VAk를 판독하고 모든 회답문서를 클러스터에 분류하여 클러스터 기억부(12)에 클러스터를 기입한다. 클러스터 분류의 순서는 제 1 실시예와 같다. 제 1 실시예와의 차이는 회답의 특징벡터 VAk를 이용하여 클러스터를 분류하는 점이다. 클러스터 라벨 작성부(17) 및 문서라벨 작성부(18)의 각각의 동작은 제 1 실시예와 같다.

이상의 동작에 의해 문서등록시에 질문과 회답에 대하여 각각 특징벡터를 추출하고, 또 회답에 대하여 클러스터, 클러스터라벨 및 문서라벨을 작성하여 각각의 기억부에 기억시킨다.

(문서검색시의 동작)

우선 인터페이스부(20)는 이용자 입력부(21)를 통해 자유문에 의한 이용자질문 Q를 접수한다. 특징벡터 추출부(15)는 이용자질문의 특징벡터 VQ를 추출한다.

도 19는 도 14 중의 특징벡터 추출부(15)에서의 이용자질문의 특징벡터 추출순서를 도시한다. 단계 S501에서는 출현하는 단어 Tj를 이용자질문 Q에서 추출하고 단어 Tj의 문서 내에서의 출현횟수 Fij를 카운트한다. 단어의 추출방법은 제 1 실시예와 같다. 단계 S502에서는 단어 Tj의 IDF값을 산출한다. 단어 Tj가 문서기억부(11) 중 어떤 문서 내에 있는 경우는 그 IDF값이 문서등록시에 이미 산출되어 있으므로 그것을 단계 S502에서 이용한다. 단어 Tj가 존재하지 않는 경우는 수학식 5에 의해 단어 Tj의 IDF값(IDFj)을 산출한다.

$$IDFj = \log(n+1) + 1$$

단계 S503에서는 이용자질문 Q에서의 단어 Tj의 가중값 WQj(TFIDF값)를 산출한다. TFIDF값의 산출방법은 제 1 실시예와 같다. 도 20은 이용자질문 Q에서 추출된 특징벡터 VQ의 예를 도시한다.

이어서 유사도 연산부(32)는 특징벡터 기억부(31)로부터 모든 질문의 특징벡터 VQi를 인출하여 이들의 특징벡터 VQi와 이용자질문의 특징벡터 VQ의 유사도를 산출한다.

도 21은 도 14 중의 유사도 연산부(32)의 처리순서를 도시한다. 단계 S511에서는 문서의 카운터 i에 i=1을 설정한다. 단계 S512에서는 특징벡터 VQi와 이용자로부터의 질문의 특징벡터 VQ의 유사도 Ei를 수학식 6에 의해 벡터의 내적으로 산출한다.

$$Ei = VQi \cdot VQ = \frac{\sum_{j} Wij \cdot WQj}{|VQi| \cdot |VQ|}$$

종료판정 단계 S513에서는 전체질문에 대해 단계 S512의 처리가 종료된 경우, 즉 i=n의 경우에는 S515로 진행한다. 그렇지 않은 경우에는 S514로 진행한다. 단계 S514에서는 카운터 i를 1 증가하여 단계 S512로 진행한다. 단계 S515에서는 모든 질문문서를 단계 S512에서 구한 유사도 Ei가 높은 순서로 분류한다.

이어서 데이터베이스 검색갱신부(33)는 유사도 연산부(32)가 산출한 유사도 Ei가 상위의 소정의 수의 질문문서와 그것에 대응하는 회답문서를 문서기억부(11)에서 검색하고 그 검색된 클러스터에 포함되는 회답문서를 다시 문서기억부(11)에서 검색하여, 그 결과를 클러스터라벨 및 문서라벨과 함께 인터페이스부(20)로 보낸다. 또 특징벡터의 유사도 연산방법으로서 벡터의 내적을 이용하였으나 벡터의 유사비를 이용하는 등 다른 방법이어도 된다.

다음으로 인터페이스부(20)는 전문가 표시부(42)를 통해 검색결과의 회답부분을 전문가에게 제시하고(도 17), 전문가 입력부(41)를 통해 전문가 표시부(42)의 표시를 참조한 전문가의 회답선택 또는 자유문에 의한 회답의 입력을 접수한다. 또 인터페이스부(20)는 이용자 표시부(22)를 통해 전문가회답을 이용자에게 제시한다(도 18). 따라서 이용자에게는 유용한 정보만이 제시된다.

도 22는 도 14 중의 데이터베이스 검색갱신부(33)의 처리순서를 플로우차트 형식으로 도시한다. 단계 S601에서는 회답사례 검색표시를 한다. 구체적으로, 인터페이스부(20)는 자유문에 의한 이용자질문 Q를 접수하고 전문가 표시부(42)를 통해 검색결과를 전문가에게 제시한다(도 17). 단계 S602에서는 검색결과를 판단한다. 전문가는 도 17의 표시를 보고 이용자질문 Q에 대하여 적절하다고 생각되는 회답이 있는지의 여부를 판단한다. 적절하다고 생각되는 회답이 있는 경우에는 S603으로 진행한다. 적절하다고 생각되는 회답이 없는 경우에는 S606으로 진행한다. 단계 S603에서, 전문가는 이용자질문 Q에 대하여 가장 적절하다고 생각되는 회답의 문서 ID를 선택한다. 인터페이스부(20)는 전문가 입력부(41)를 통해 선택된 문서 ID의 입력을 접수한다. 또 당해 문서 ID를 후술하는 단계 S605를 위해 데이터베이스 검색갱신부(33)에 주고받는다. 단계 S604에서 인터페이스부(20)는 이용자 표시부(22)를 통해 전문가가 선택한 문서 ID의 문서를 회답으로서 이용자에게 제시한다(도 18).

단계 S605에서는 질문추가처리를 한다. 데이터베이스 검색갱신부(33)는 주고받은 문서 ID의 회답에 대응하는 1 이상의 질문중 이용자질문 Q와의 유사도가 가장 높은 질문의 유사도가 소정의 값 이하인 경우에는 적절한 자동회답이 이루어지지 않은 것으로 하여, 도 15의 질문표에 신규의 독특한 질문 ID, 이용자질문 Q 및 선택된 문서 ID로 이루어지는 행을 추가한다. 이어서 단계 S612로 진행한다. 단계 S612에서는 특징벡터 추출부(15)는 문서등록시와 마찬가지로 문서기억부(11)에 기억된 모든 질문 Qi 및 회답 Ak로부터 각각의 특징벡터 VQi, VAk를 추출하여 추출된 특징벡터를 특징벡터 기억부(31)에 기입한다.

단계 S602에서 적절한 회답이 없는 경우, 전문가는 단계 S606에서 이용자질문 Q에 대하여 적절한 회답 A를 자유문으로 입력한다. 인터페이스부(20)는 전문가입력부(41)를 통해 자유문의 회답을 접수한다. 단계 S607에서 인터페이스부(20)는 전문가가 입력한 회답 A를 이용자에게 제시한다. 단계 S608에서 특징벡터 추출부(15)는 전문가가 입력한 회답 A의 특징벡터 VA를 추출한다. 이 특징벡터의 추출순서는 도 19에서 설명한 이용자질문 Q의 특징벡터 VQ의 추출순서와 같다. 단계 S609에서, 유사도 연산부(32)는 특징벡터 기억부(31)로부터 모든 회답의 특징벡터 VAk를 인출하여 전문가가 입력한 회답 A의 특징벡터 VA와의 유사도 Ek를 산출한다. 이 유사도의 산출순서는 도 21에서 설명한 이용자질문 Q의 유사도의 산출순서와 같다. 단계 S610에서 유사도 연산부(32)는 단계 S609에서 구한 유사도 Ek 중에서 가장 큰 것이 소정의 값 이상인 경우는 문서기억부(11) 내에 전문가가 입력한 회답 A와 유사한 회답이 있는 것으로 판단하여 유사한 회답 Ak의 문서 ID를 데이터베이스 검색갱신부(33)에 교환하고 단계 S605로 진행한다. 그렇지 않은 경우는 단계 S611로 진행한다. 단계 S611에서는 질문회답 추가처리를 한다. 데이터베이스 검색갱신부(33)는 도 16의 회답표에 신규의 독특한 문서 ID 및 전문가가 입력한 회답 A로 이루어지는 행을 추가한다. 또 도 15의 질문표에 신규의 독특한 질문 ID, 이용자질문 Q 및 추가한 회답에 부여한 문서 ID로 이루어지는 행을 추가한다. 그리고 단계 S612로 진행한다. 단계 S612에서의 처리는 상술한 바와 같다.

한편 회답을 선택 또는 입력할 수 있는 전문가가 없는 경우에 인터페이스부(20)는 이용자 표시부(22)를 통해 도 17과 같은 검색결과를 이용자에게 제시한다. 이용자는 도 17의 표시를 보고 자기의 질문 Q에 대하여 가장 적절하다고 생각되는 회답의 문서 ID를 선택하고, 인터페이스부(20)는 이용자 입력부(21)를 통해 선택된 문서 ID의 입력을 접수한다. 데이터베이스 검색갱신부(33)는 입력된 문서 ID의 회답에 대응하는 1 이상의 질문중 이용자질문 Q와의 유사도가 가장 높은 질문의 유사도가 소정의 값 이하인 경우에는 적절한 자동회답이 이루어지지 않은 것으로 하여 도 15의 질문표에 신규의 독특한 질문 ID, 이용자질문 Q 및 선택된 문서 ID로 이루어지는 행을 추가한다(단계 S605와 같음). 그리고 특징벡터 추출부(15)는 문서등록시와 마찬가지로 문서기억부(11)에 기억된 모든 질문 Qi 및 회답 Ak에서 각각의 특징벡터 VQi, VAk를 추출하여 추출된 특징벡터를 특징벡터 기억부(31)에 기입한다(단계 S612와 같음).

이상과 같이 제 2 실시예에 의하면 이용자 또는 전문가의 응답에 따라 문서기억부(11)가 자동적으로 갱신되도록 하였으므로 다음에 같은 질문이 입력된 경우에 적절한 회답을 할 수 있는 정보검색 시스템을 제공할 수 있다.

발명의 효과

이상 설명한 바와 같이 본 발명에 의하면 문서의 특징벡터를 산출하여 특징벡터에 기초하여 문서를 클러

스터 분류하고 문서의 검색결과를 클러스터별로 정리하고 표시하였으므로 이용자는 검색결과를 유사한 문서의 집합으로서 쉽게 파악할 수 있게 된다. 따라서 이용자의 정보검색에 소요되는 부담을 경감시키는 정보검색 시스템을 제공할 수 있다.

또 본 발명에 의하면 이용자로부터 질문이 입력된 경우에 유사질문을 검색하고 대응하는 회답을 이용자 또는 전문가에게 제시하여 이용자 또는 전문가가 가장 적절하다고 판단한 회답을 선택하였을 때에는 그 선택된 회답을 기초로 또 적절한 회답이 없을 때에는 전문가 새롭게 입력한 회답을 기초로 문서 데이터베이스를 자동적으로 갱신하는 것이므로 검색대상의 정보를 용이하게 갱신할 수 있는 정보검색 시스템을 제공할 수 있다.

(57) 청구의 범위

청구항 1

복수의 문서 중에서 이용자가 구하는 정보를 검색하기 위한 정보검색 시스템에 있어서,

상기 복수의 문서를 기억하기 위한 문서기억수단과,

상기 문서기억수단에 기억된 복수의 문서의 각각의 특징량을 추출하기 위한 특징량 추출수단과,

상기 추출된 특징량에 기초하여 상기 문서기억수단에 기억된 복수의 문서를 각 클러스터가 하나의 문서 또는 서로 근사한 특징량을 갖는 복수의 문서로 이루어지도록 복수의 클러스터로 분류하기 위한 클러스터 분류수단과,

상기 문서기억수단에 기억된 복수의 문서 중에서 상기 이용자로부터 주어진 검색조건을 만족하는 문서를 검색하기 위한 문서검색수단과,

상기 검색된 문서를 당해 문서가 속하는 클러스터가 복수의 문서로 이루어지는 경우에 당해 클러스터 중 다른 문서와 함께 검색결과로서 제시하기 위한 인터페이스수단을 구비하는 것을 특징으로 하는 정보검색 시스템.

청구항 2

제 1항에 있어서,

상기 특징량 추출수단은 상기 문서기억수단에 기억된 복수의 문서의 각각으로부터 당해 문서 중에 출현하는 1 또는 복수의 단어와 당해 단어가 당해 문서를 특징짓는 가중값과의 그룹을 요소로 하는 특징벡터를 상기 특징량으로서 추출하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 3

제 1항에 있어서,

상기 클러스터 분류수단은 복수의 문서로 된 클러스터의 수가 최대가 되는 클러스터링을 채용하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 4

제 1항에 있어서,

각각 상기 복수의 클러스터중 대응하는 클러스터의 내용을 나타내는 복수의 클러스터라벨을 작성하기 위한 클러스터라벨 작성수단을 추가로 구비하며,

상기 인터페이스수단은 상기 작성된 복수의 클러스터 라벨중 상기 검색된 문서가 속하는 클러스터의 내용을 나타내는 클러스터라벨을 상기 검색결과와 함께 제시하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 5

제 4항에 있어서,

상기 클러스터라벨 작성수단은 상기 복수의 클러스터의 각각에 대하여 당해 클러스터에 속하는 모든 문서 중에서 당해 클러스터를 특징짓는 1 또는 복수의 단어를 상기 클러스터 라벨로서 선택하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 6

제 4항에 있어서,

상기 클러스터라벨 작성수단은 상기 복수의 클러스터의 각각에 대하여 당해 클러스터에 속하는 모든 문서 중에서 당해 클러스터를 특징짓는 하나의 문장을 상기 클러스터 라벨로서 선택하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 7

제 4항에 있어서,

각각 상기 문서기억수단에 기억된 복수의 문서중 대응하는 문서의 내용을 나타내는 복수의 문서라벨을 작성하기 위한 문서라벨 작성수단을 추가로 구비하며,

상기 인터페이스수단은 상기 작성된 복수의 문서라벨중 상기 검색된 문서가 속하는 클러스터 중의 각 문서의 내용을 나타내는 문서 라벨을 상기 검색결과와 함께 제시하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 8

제 7항에 있어서,

상기 문서라벨 작성수단은 상기 문서기억수단에 기억된 복수의 문서의 각각에 대하여 당해 문서 중의 모든 문서 중에서 당해 문서를 특징짓는 하나의 문장을 상기 문서라벨로서 선택하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 9

제 1항에 있어서,

상기 복수의 문서는 서로 대응된 복수의 질문문서와 복수의 회답문서를 포함하며,

상기 검색조건은 자유문에 의한 이용자 질문이고,

상기 특징량 추출수단은 상기 문서기억수단에 기억된 복수의 회답문서가 상기 클러스터 분류수단에 의해 복수의 클러스터에 분류되도록 상기 문서기억수단에 기억된 복수의 회답문서의 각각의 특징량을 추출하고,

상기 정보검색 시스템은 상기 문서기억수단에 기억된 복수의 질문문서의 각각에 대하여 상기 이용자 질문에 관한 문서와의 사이의 유사도를 산출하기 위한 유사도 연산수단을 추가로 구비하며,

상기 문서검색수단은 상기 산출된 유사도에 기초하여 상기 문서기억수단에 기억된 복수의 질문문서 중에서 유사도가 높은 질문문서를 검색하고, 상기 문서기억수단에 기억된 복수의 회답문서 중에서 상기 검색된 질문문서에 대응된 회답문서를 검색하며,

상기 인터페이스수단은 상기 검색된 회답문서를 그 검색된 회답문서가 속하는 클러스터가 복수의 회답문서로 이루어지는 경우에 당해 클러스터중 다른 회답문서와 함께 상기 검색결과로서 제시하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 10

제 9항에 있어서,

상기 인터페이스수단은 상기 검색결과를 상기 이용자에게 제시하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 11

제 10항에 있어서,

상기 인터페이스수단은 상기 제시된 검색결과 중에서 상기 이용자에 의한 회답문서의 선택을 접수하도록 구성되고,

상기 정보검색 시스템은 상기 문서기억수단에 기억된 복수의 질문문서 중에서 상기 선택된 회답문서에 대응된 질문문서를 검색하여 그 검색된 질문문서와 상기 이용자질문에 관한 문서와의 유사도가 소정의 값보다 낮은 경우에 상기 이용자질문에 관한 문서를 상기 선택된 회답문서와 대응하여 상기 문서기억수단에 새롭게 기억시키기 위한 문서갱신수단을 추가로 구비하는 것을 특징으로 하는 정보검색 시스템.

청구항 12

제 9항에 있어서,

상기 인터페이스수단은 상기 검색결과를 상기 이용자질문에 관한 문서와 함께 전문가에게 제시하고, 당해 제시된 검색결과 중에서 상기 전문가에 의해 선택된 회답문서를 상기 이용자에게 제시하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 13

제 12항에 있어서,

상기 문서기억수단에 기억된 복수의 질문문서 중에서 상기 선택된 회답문서에 대응된 질문문서를 검색하고 당해 검색된 질문문서와 상기 이용자질문에 관한 문서와의 유사도가 소정의 값보다 낮은 경우에는 상기 이용자질문에 관한 문서를 상기 선택된 회답문서와 대응하여 상기 문서기억수단에 새롭게 기억시키기 위한 문서갱신수단을 추가로 구비하는 것을 특징으로 하는 정보검색 시스템.

청구항 14

제 9항에 있어서,

상기 인터페이스수단은 상기 검색결과를 상기 이용자질문에 관한 문서와 함께 전문가에게 제시하고, 당해 제시된 검색결과를 참조하여 상기 전문가가 자유문으로 입력한 회답문서를 상기 이용자에게 제시하도록 구성된 것을 특징으로 하는 정보검색 시스템.

청구항 15

제 14항에 있어서,

상기 문서기억수단에 기억된 복수의 회답문서의 각각과 상기 입력된 회답문서와의 유사도가 어느 것이나 소정의 값보다 낮은 경우에는 상기 이용자질문에 관한 문서와 상기 입력된 회답문서를 서로 대응하여 상기 문서기억수단에 새롭게 기억시키기 위한 문서갱신수단을 추가로 구비하는 것을 특징으로 하는 정보검색 시스템.

청구항 16

복수의 문서 중에서 이용자가 구하는 정보를 검색하기 위한 정보검색 시스템에 있어서,

서로 대응된 복수의 질문문서와 복수의 회답문서를 기억하기 위한 문서기억수단과,

상기 이용자로부터 자유문에 의한 이용자질문이 주어졌을 때 상기 문서기억수단에 기억된 복수의 질문문서의 각각에 대하여 상기 이용자질문에 관한 문서와의 사이의 유사도를 산출하기 위한 유사도 연산수단과,

상기 산출된 유사도에 기초하여 상기 문서기억수단에 기억된 복수의 질문문서 중에서 유사도가 높은 복수의 질문문서를 검색하고, 상기 문서기억수단에 기억된 복수의 회답문서 중에서 상기 검색된 복수의 질문문서의 각각에 대응된 회답문서를 검색하기 위한 문서검색수단과,

상기 이용자질문에 관한 문서와 함께 상기 검색된 복수의 회답문서를 검색결과로서 전문가에게 제시하고, 당해 제시된 검색결과 중에서 상기 전문가에 의해 선택된 회답문서 또는 당해 제시된 검색결과를 참조하여 상기 전문가가 자유문으로 입력한 회답문서를 상기 이용자에게 제시하기 위한 인터페이스수단을 구비하는 것을 특징으로 하는 정보검색 시스템.

청구항 17

제 16항에 있어서,

상기 문서기억수단에 기억된 복수의 질문문서 중에서 상기 선택된 회답문서에 대응된 질문문서를 검색하고, 당해 검색된 질문문서와 상기 이용자질문에 관한 문서와의 유사도가 소정의 값보다 낮은 경우에 상기 이용자 질문에 관한 문서를 상기 선택된 회답문서와 대응하여 상기 문서기억수단에 새롭게 기억시키기 위한 문서갱신수단을 추가로 구비하는 것을 특징으로 하는 정보검색 시스템.

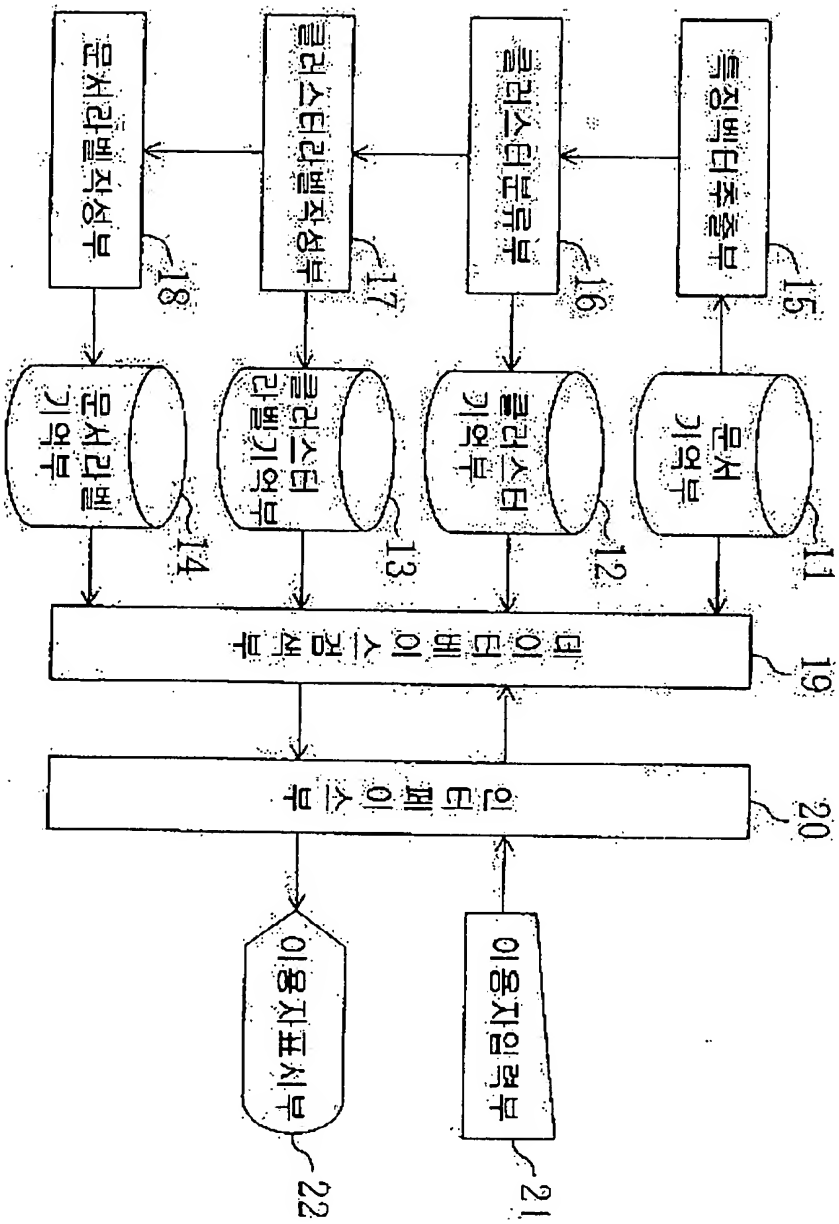
청구항 18

제 16항에 있어서,

상기 문서기억수단에 기억된 복수의 회답문서의 각각과 상기 입력된 회답문서와의 유사도가 어느 것이나 소정의 값보다 낮은 경우에는 상기 이용자질문에 관한 문서와 상기 입력된 회답문서를 서로 대응하여 상기 문서기억수단에 새롭게 기억시키기 위한 문서갱신수단을 추가로 구비하는 것을 특징으로 하는 정보검색 시스템.

도면

도면1


특징벡터추출부
클러스터분류부
클러스터라벨작성부
문서라벨작성부
15
16
17
18
문서 기억부
클러스터 기억부
클러스터라벨기억부
문서라벨 기억부
11
12
13
14
데이터베이스부
19
인터페이스부
20
이용자입력부
이용자표시부
21
22

도면2

문서ID
본문
문서ID:
1
과자가 먹고싶어졌을때는
① 씹는 효과가 있고 미련이 남지 않는 것으로 양을 정하여 먹는다. ...
2
운동은 언제 어디에서나 혼자하는 것이 원칙입니다.
사람에 따라 생활, 라디오이 모두 다릅니다. ...
3
완전히 당신을 생각과 같습니다. 무언가 고민이 있거나, ...
: 이미

도면3

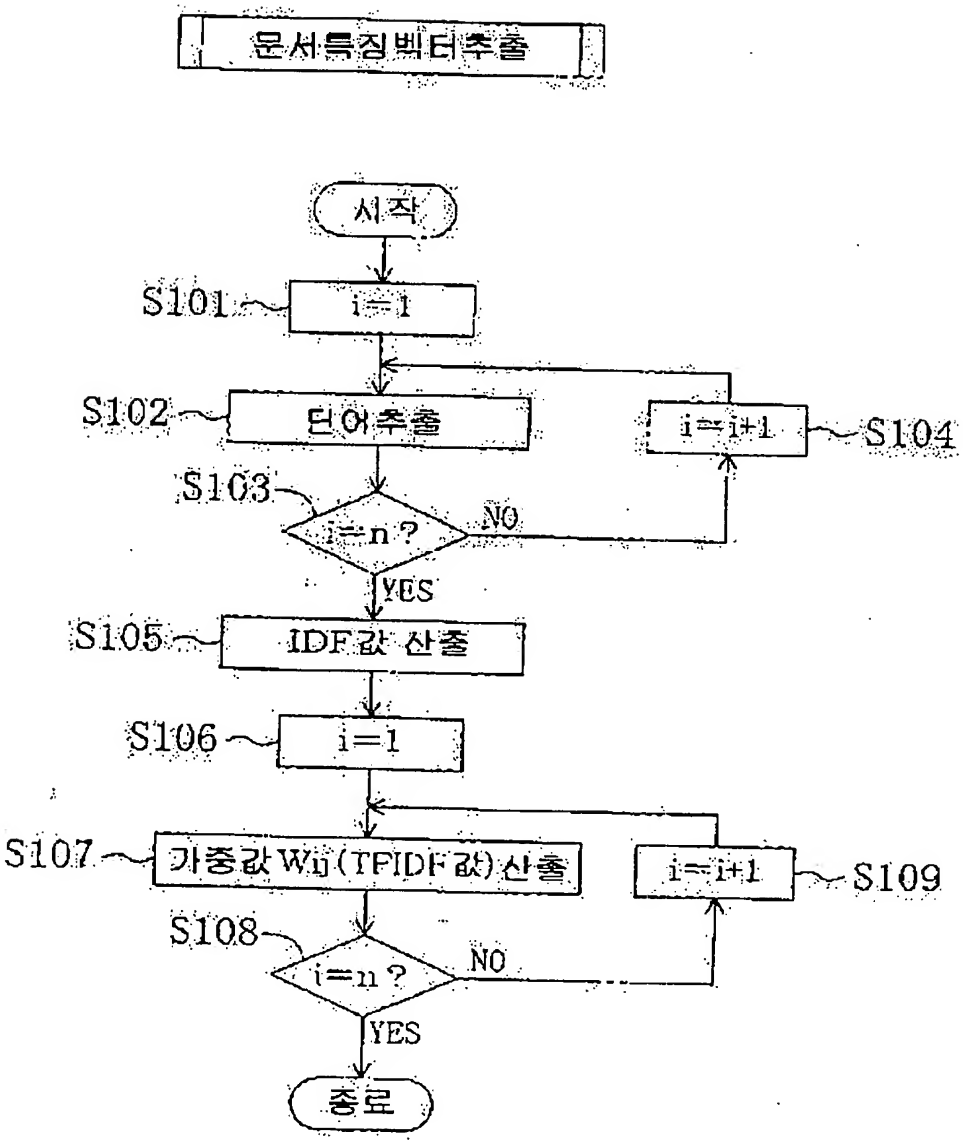
OO건의 검색결과가 있었습니다.

클러스터ID	클러스터타이틀	문서ID	문서
1	과자 간식 치즈	1	과자가 먹고싶어졌을때는, 오늘는 요과가있고 미혈이 남자 않는 컷으로 앞을 정하여 먹는다. ...
		190	간식은 우유, 유제품(치즈, 요구르트 등), ...
	수분이 많은 것(젤리, 푸딩, 요구르트)을...	432	간식은 200kcal 이내에서 자유롭게 선택해도 됩니다. ... 식량로리로 하고 싶은 음우는 식량로리 심니표의...
		644	과자가 먹고싶어졌을때는, ·씹는 효과기 있고 미혈이 남자 않는 것으로,... 어떠 경우에도 1일의 토달이 200kcal 이내이고, ...

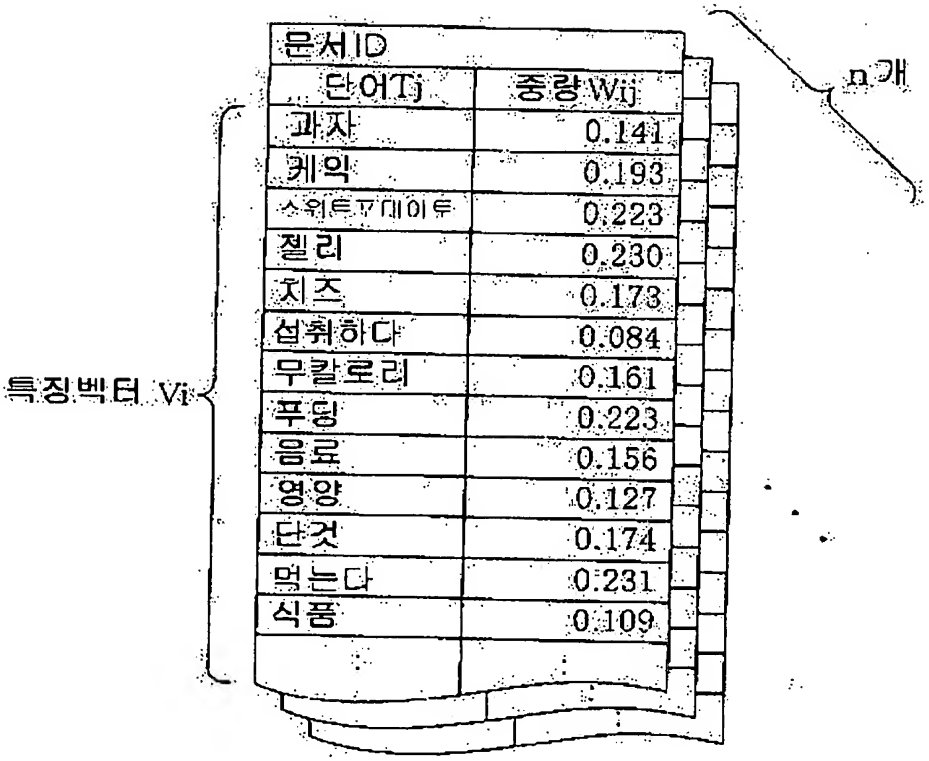
이전 클러스터

다음 클러스터

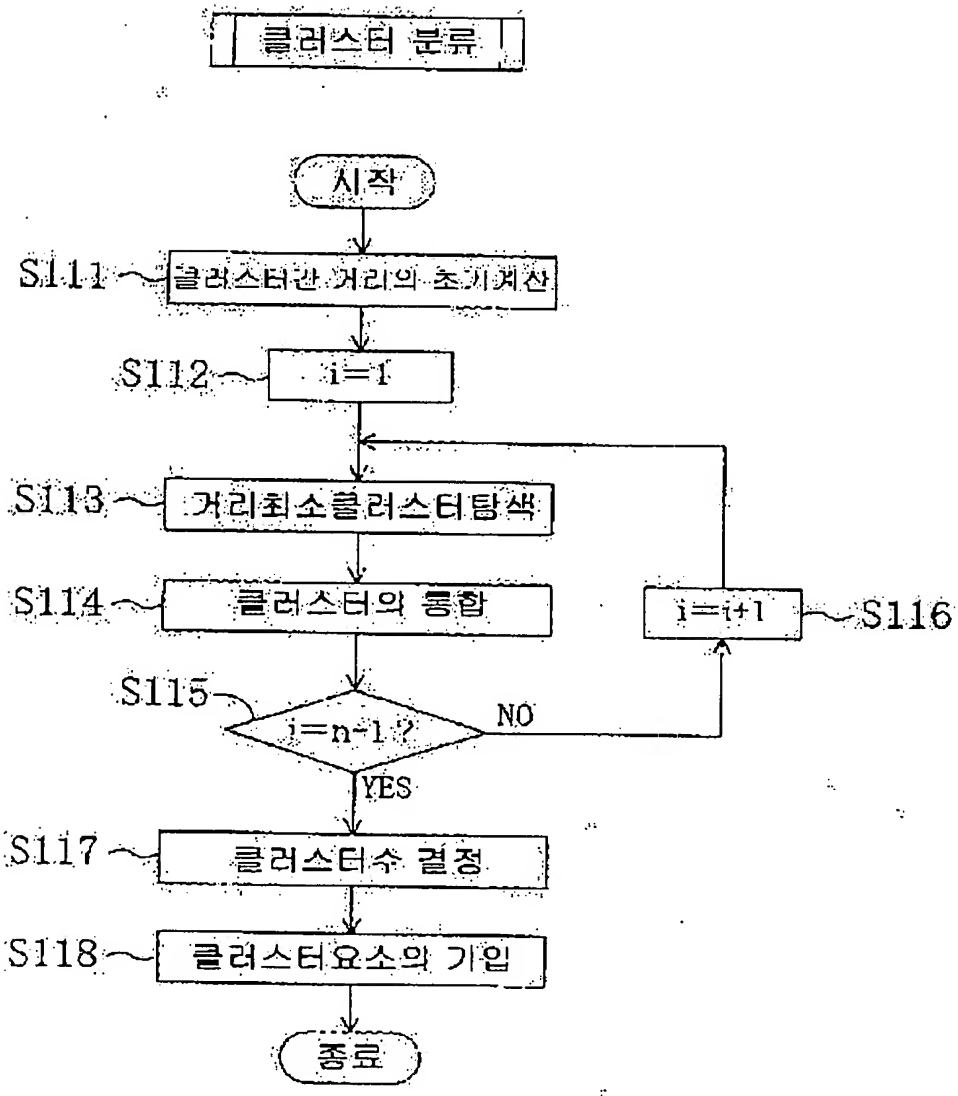
도면4


문서특징벡터추출
시작
S101
i=1
S102
단어추출
i=i+1
S104
S103
i=n ?
NO
YES
S105
IDF값 산출
S106
i=1
S107
가중값 Wij (TFIDF값) 산출
i=i+1
S109
S108
i=n ?
NO
YES
종료

도면5


문서ID
단어Tj
중량Wij
과자 0.141
케익 0.193
스위트포테이토 0.223
젤리 0.230
치즈 0.173
섭취하다 0.084
무칼로리 0.161
푸딩 0.223
음료 0.156
영양 0.127
단것 0.174
먹는다 0.231
식품 0.109
n개
특징벡터 Vi

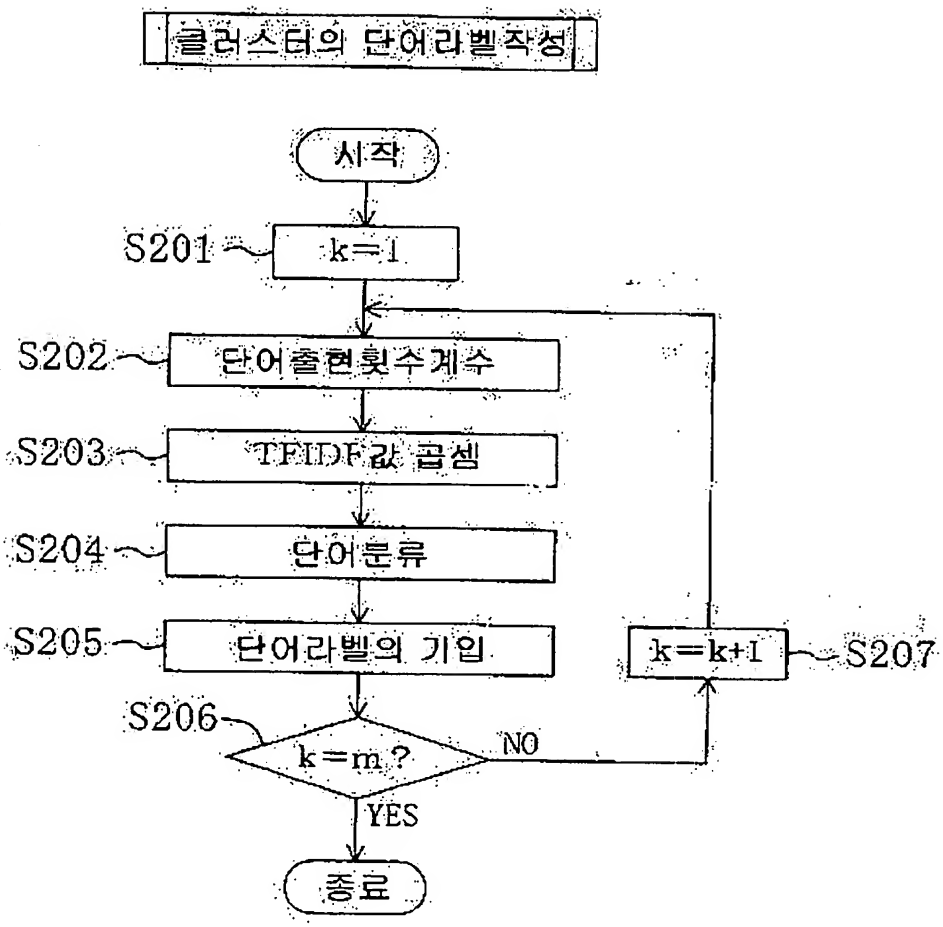
도면6



도면7

클러스터ID	문서ID
1	1, 190, 432, 644
2	2, 412, 3, 158
3	3, 158
4	4, 109, 182, 615
:	:

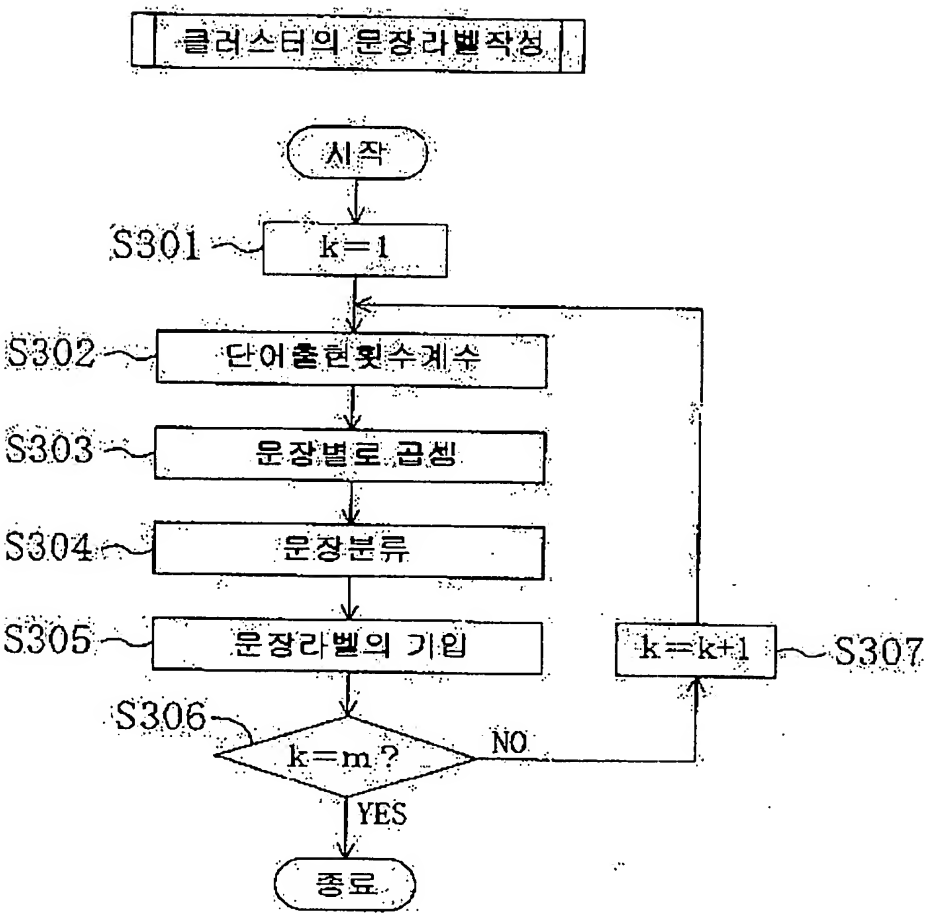
도면8



도면9

클러스터ID	단어라벨
1	과자, 간식, 치즈
2	컨디션, 연습, 효과
3	스트레스, 적극적, 상황
4	생리, 식욕, 다이어트
:	:

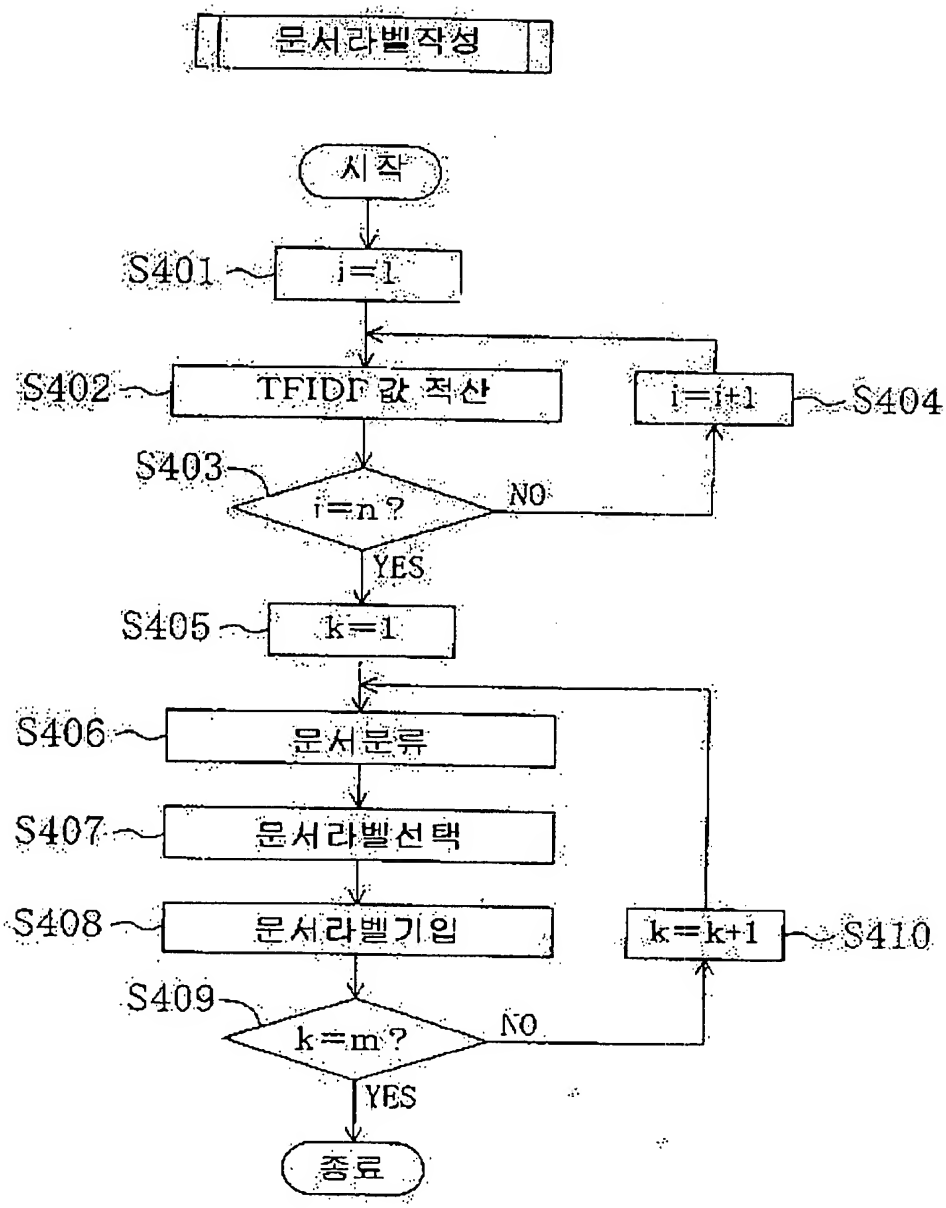
도면10



도면11

클러스터ID	문장라벨
1	수분이 많은 것(젤리, 푸딩, 요구르트)을 ...
2	생리중, 컨디션이 나쁠때는 무리하게 운동을 ...
3	무엇이 스트레스가 되었는지를 확인하여, ...
4	따라서 생리전이라도 자신의 다이어트 페이스를 유지하도록 ...
:	:

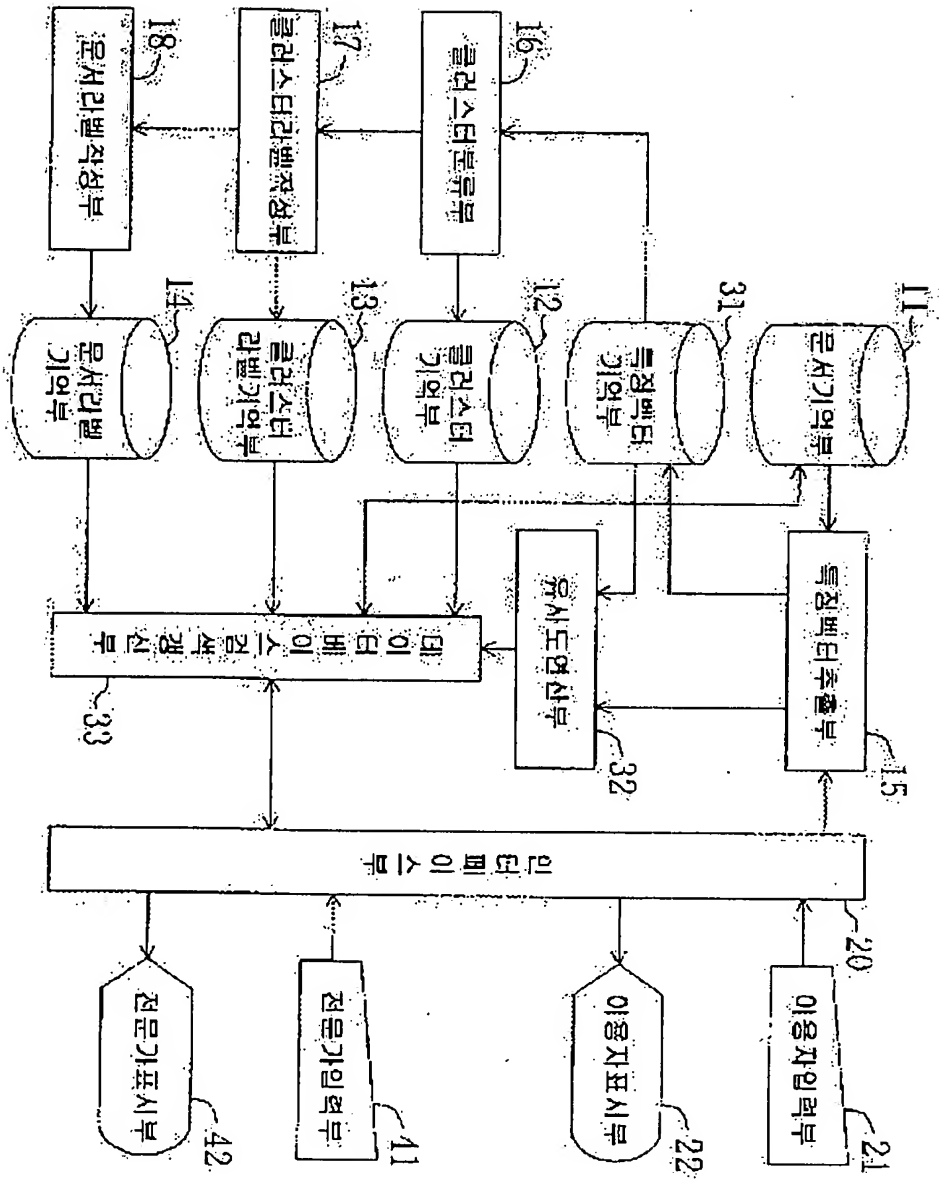
도면12


문서라벨작성
시작
S401
i=1
S402
TFIDF값 적산
i=i+1
S404
S403
i=n?
NO
YES
S405
k=1
S406
문서분류
S407
문서라벨선택
S408
문서라벨기입
k=k+1
S410
S409
k=m?
NO
YES
종료

도면13

클러스터ID	문서ID	문서라벨
1	1	썰은 홍과가 있고 미련이 남지 않는 건으로 ...
	190	간식에는 우유, 유제품(치즈, 요구르트 등) ...
	432	저칼로리로 하고 싶은 경우에는 저칼로리 감미료의 ...
	644	이 때 경우에도 1일의 도달이 200kcal 이내이고, ...
...	...	...

도면14


문서기억부
11
특정벡터 기억부
31
클러스터 기억부
12
클러스터 리벨기억부
13
문서리벨 기억부
14
특정벡터추출부
15
유사도연산부
32
데이터베이스 관리부
33
입력부
20
이용자입력부
21
이용자표시부
22
전문가입력부
41
전문가표시부
42
클러스터분류부
16
클러스터리벨작성부
17
문서리벨작성부
18

도면15

질문표

질문ID	질문	대응 문서ID
1	아무래도 과자가 먹고싶다면 어떤 것을 먹으면 좋겠습니까? ···	1
2	아침식사전의 운동이 가장 효과가 있다고 들었습니다만, 저녁의 유산소운동은 어떨 습니까? ···	2
3	이번달의 생리가 1주일 이상이나 늦어졌습니다. ···	3
⋮	⋮	⋮

n개

도면16

회답표

문서ID	회답
1	과자가 먹고싶을때는, ①씹는 효과가 있고 미련이 남지 않는 것으로, 양을 정하여 먹는다.···
2	운동은 언제 어디서나 혼자 하는 것이 원칙입니다. 사람마다 생활, 컨디션은 모두 다릅니다. ···
3	완전히 당신의 생각과 같습니다. 무언가 고민이 있거나, ···
⋮	⋮

m개

도면17

질문
아무래도 과자가 먹고싶어졌다면 어떤 것을 먹으면 좋겠습니까. ...

회답			
문서ID	문서	문서리뷸	리스트ID
1	과자가 먹고 싶어졌을 때는 (1)별로 칼로리가 있고, 미량이 남지 않는 것으로 양을 정하여 먹는다. ...		
190	간식에는 우유, 유제품(치즈, 요구르트 등), ...	수분이 많은 것(젤리, 푸딩, 요구르트 등), ...	1
432	간식은 200kcal 이내에서 자유롭게 선택해도 됩니다. ... 저칼로리로 하고 싶은 경우는 저칼로리 강미르의 ...		
644	과자가 먹고 싶어졌을 때는 , 썹는 효과가 있고, 미량이 남지 않는 것으로... 모두 1일의 토탈이 200kcal 이내에서, ...		

이전 클리스터

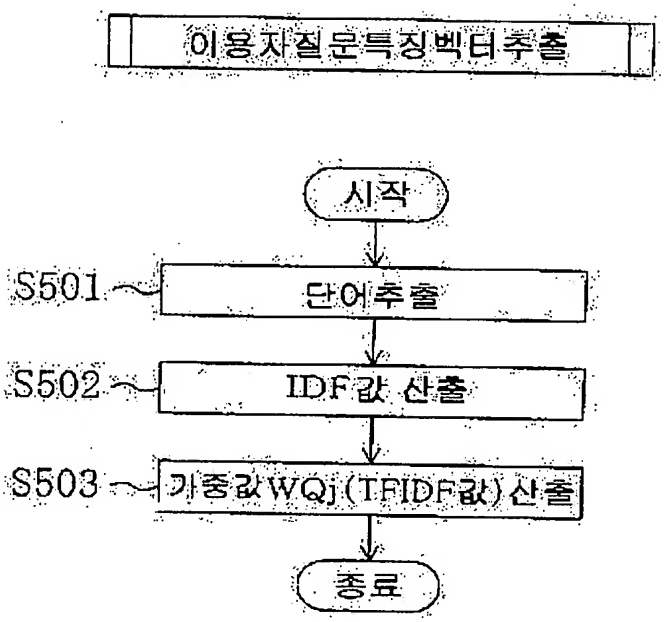
다음 클리스터

도면18

질문
아무래도 피자가 먹고 싶으면 어떤 것을 먹으면 좋겠습니까? ...

답
피자가 먹고 싶어졌을 때는, ⓐ씹는 효과가 있고, 미러이닝지 않는 것으로 양을 천하여 먹는다. ...

도면19

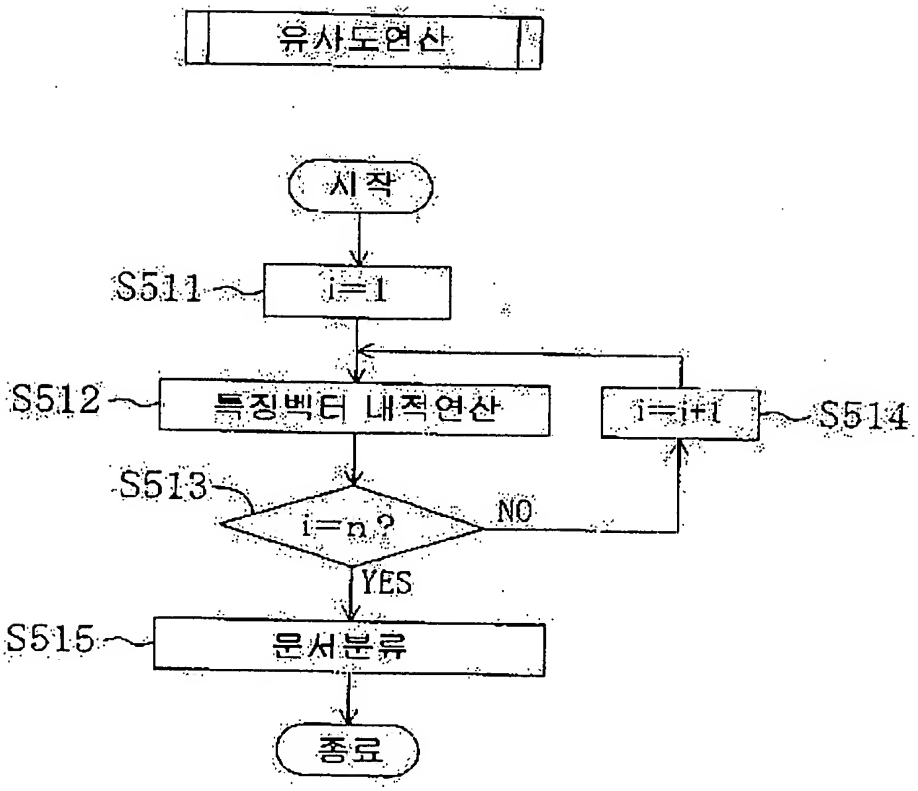


도면20

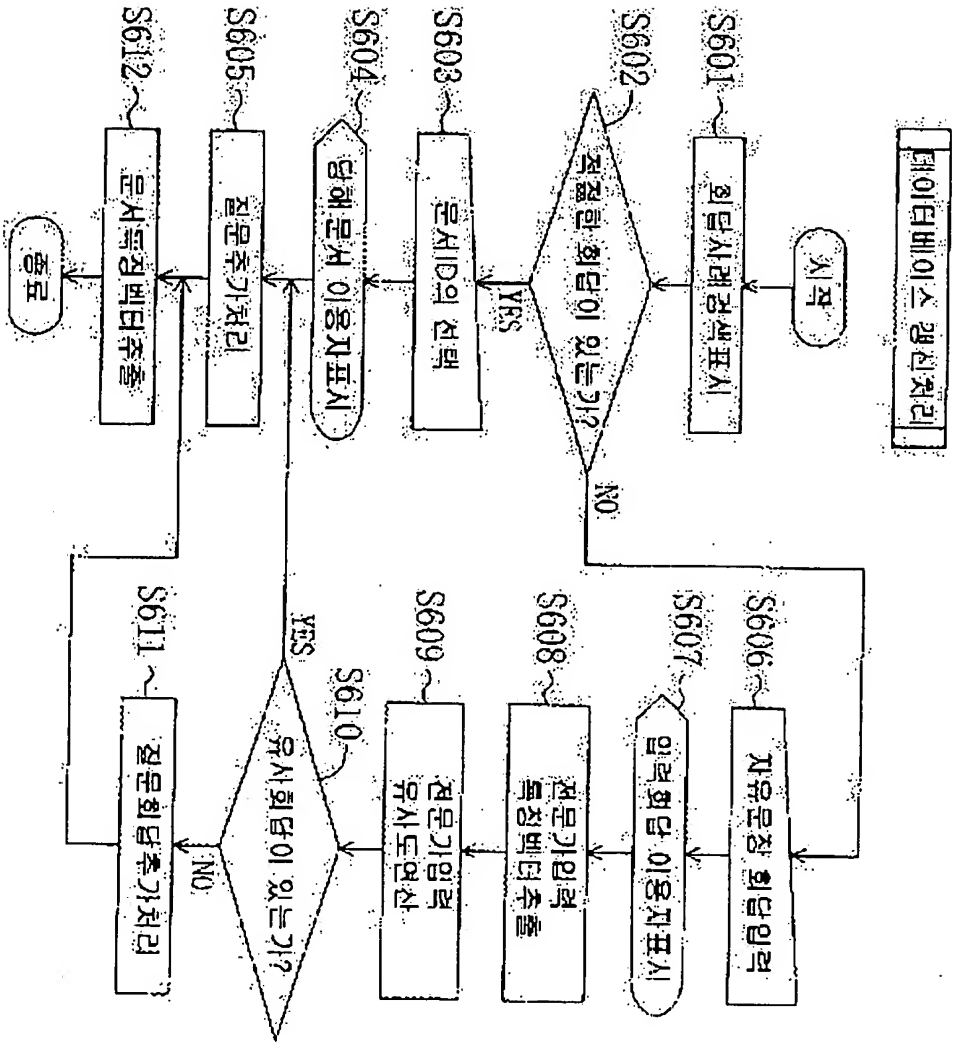
특징벡터VQ

단어Tj	중량WQj
과자	0.601
간식	0.452
나	0.400
먹는다	0.847
아침식사	0.556
방법	0.456
:	:

도면21


유사도연산
시작
S511
i=1
S512
특징벡터 내적연산
i=i+1
S514
S513
i=n ?
NO
YES
S515
문서분류
종료

도면22


데이터베이스 갱신처리
시작
S601
회답사례검색표시
S602
적절한 회답이 있는가?
YES
NO
S603
문서ID의 선택
S604
당해문서 이용자표시
S605
질문추가처리
S612
문서특정벡터추출
종료
S606
자유문장 회답입력
S607
입력회답 이용자표시
S608
전문기입력 특징벡터추출
S609
전문기입력 유사도연산
S610
유사회답이 있는가?
YES
NO
S611
질문회답추가처리